Tailored IoT & BigData Sandboxes and Testbeds for Smart, Autonomous and Personalized Services in the European Finance and Insurance Services Ecosystem

D3.13 – Data Governance Frameworks and Tools II

Revision Number	3.0
Task Reference	T3.5
Lead Beneficiary	GRAD
Responsible	Inés Ortega Fernández
Partners	GRAD, ATOS, JSI
Deliverable Type	Report (R)
Dissemination Level	Public (PU)
Due Date	2021-07-31
Delivered Date	2021-07-28
Internal Reviewers	UBI, BOUN
Quality Assurance	INNOV
Acceptance	WP Leader Accepted and/or Coordinator Accepted
EC Project Officer	Pierre-Paul Sondag
Programme	HORIZON 2020 - ICT-11-2018
	This project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement no 856632

Contributing Partners

Partner Acronym	**Role**[1]	**Author(s)**[2]
GRAD	Lead Beneficiary	Inés Ortega Fernández Lilian Adkinson Orellana Sara ElKortbi Martínez
ATOS	Contributor	Nuria Ituarte Aranda
JSI	Contributor	Maja Škrjanc Klemen Kenda Beno Šircelj
UBI	Internal Reviewer	Dimitris Miltiadou
INNOV	Quality Assurance	John Soldatos

Revision History

Version	Date	Partner(s)	Description
0.1	2021-04-28	GRAD	ToC Version
0.2	2021-06-03	ATOS	Added ATOS contribution to Section 2.3
0.3	2021-06-07	GRAD	Added GRAD contribution to Section 2.2
0.4	2021-06-17	JSI	Added JSI contribution to Section 2.1
0.5	2021-06-18	ATOS	Added ATOS contribution to Section 3.3
0.6	2021-06-39	GRAD	Added GRAD contribution to Section 3.2
0.7	2021-07-14	JSI	Added JSI contribution to Section 3.1
1.0	2021-07-15	GRAD, ATOS, JSI	Version for Internal Review
2.0	2021-07-23	GRAD	Version for Quality Assurance
3.0	2021-07-30	GRAD	Version for Submission

[1] Lead Beneficiary, Contributor, Internal Reviewer, Quality Assurance

[2] Can be left void

Executive Summary

This document exposes how Data Governance Framework and Tools are developed through the INFINITECH project to ensure appropriate behaviour of Big Data and Data Analytics frameworks. The increment of the amount of data handled in Big Data contexts constitutes a challenge for managing data privacy: the velocity, variety and complexity of the data make necessary the use of automated tools that help manage the risks and regulatory requirements associated with such contexts. In addition, management of user identities is also challenging in Digital Finance: FinTechs need to onboard users into their platforms in a secure and reliable manner, using mechanisms that replace the need of manually inspecting physical identity documents.

With regards to the General Data Protection Regulation (GDPR) [1], depending on the nature of the data different data privacy and security measures need to be implemented: pseudonymization is considered a security measure to allow the processing of personal data, while anonymization makes the individuals within a dataset non-identifiable, and therefore anonymized data is not considered personal data anymore.

Data Governance is an essential mechanism to guarantee data security and privacy, as well as to establish the required workflows to manage data and information in companies, and in particular in banks, FinTechs and other insurance and financial organisations. To meets its goals, the INFINITECH project provides the following tools and mechanisms:

a) **A pseudonymization tool;**

b) **A tool for anonymizing data**, and

c) **A mobile digital user onboarding service.**

The first tool provides mechanisms to pseudonymize unique identifiers, and the generalization of numeric and time-stamped enriched transactional data by exploiting different techniques. The second tool provides an automatic framework for anonymizing datasets by automatically selecting the best configuration which fits the aspired privacy and utility goals. Finally, the third tool, namely the Digital User Onboarding Services (DUOS) - provides a mobile service that allows the creation of virtual identities by combining digital certificates from government issued electronic IDs or passports with face images.

The deliverable documents the updates of the design specifications initially described in D3.12 “Data Governance Frameworks and Tools I” [2] to the most current version. Furthermore, a description of the prototype implementations carried out over the last period is presented. Finally, a set of conclusions is included in order to summarize the most important concepts and the topics addressed.

Table of Contents

List of Figures

List of Tables

Abbreviations/Acronyms

Abbreviation	Definition
4AMDL	Anti-Money Laundering Directive IV
AI	Artificial Intelligence
API	Application Programming Interface
ARIES	ReliAble euRopean Identity EcoSystem
BIC	Business Identifier Code
BOS	Bank of Slovenia
CLI	Command Line Interface
CSV	Comma-Separated Values
DPO	Data Protection Orchestrator
DUOS	Digital User Onboarding Service
eDNI	electronic National Identity Document
eID	electronic ID
eMRTD	Electronic Machine Readable Travel Documents
GDPR	General Data Protection Regulation
GPS	Global Positioning System
GUI	Graphical User Interface
HPC	High Performance Computing
HTTP	HyperText Transfer Protocol
IAM	Identity and Access Management
IBAN	International Bank Account Number
ICAO	International Civil Aviation Organization
ID	IDentifier
ILM	Information Loss Measure
IoT	Internet of Things
JSON	JavaScript Object Notation
MAE	Mean Absolute Error
MiFiD	Markets in Financial Instruments Directive
MD5	Message Digest Algorithm Five
ML	Machine Learning
MRZ	Machine-Readable Zone
MSE	Mean Squared Error
MV	Mean Variation
NFC	Near-Field Communication
OCR	Optical Character Recognition
PSD	Payment Service providers Directive
PAMLS	Platform for Anti Money Laundering Supervision
QR	Quick Response
RA	Reference Architecture
REST	REpresentational State Transfer

SHARP	Smart, Holistic, Autonomy, Personalized and Regulatory Compliance
SP	Service Provider
SPeID	Service Provider for eIDas Integration service
URL	Uniform Resource Locator
vID	virtual IDentifier

1 Introduction

The current document is the second deliverable of a series of three deliverables whose goal is to present the data governance mechanisms that will be developed within the context of the INFINITECH project during the 26 months of the task "T3.5 Data Governance Mechanisms". In this deliverable, we provide the updates of the design of the data governance tools and we present the technical advances during this period.

1.1 Objective of the Deliverable

The objective of this deliverable is twofold: on the one hand, it includes an updated design of each of the data governance mechanisms developed within task "T3.5 Data Governance Mechanisms"; on the other hand, it presents the technical advances on the development of each tool, either by describing them at a technical or at conceptual level. Specifically, the mechanisms are the following:

1. A **pseudonymization tool** to pseudonymize enriched transactional data.
2. A **tool for anonymizing datasets** that determines automatically the best anonymization configuration for a particular dataset.
3. A **mobile digital user onboarding service** which uses virtual IDs derived from government issued documents (such as eID cards or passports).

Deliverable D3.12 "Data Governance Framework and Tools I" [2] included the definition of a **Service Provider for eIDas Integration service (SPeID)**. SPeID is a solution for authenticating citizens against pan European eIDAS infrastructure, providing a strong cross-border authentication mechanism based on eIDs, and ensuring that cross-border transactions are legal. This component is not going to be continued within the INFINITECH project, and therefore **it is not included in the present deliverable.** As there is no requirement from the INFINITECH pilots to include cross-border authentication functionality, and therefore **its implementation cannot be validated within INFINITECH.** Moreover, eIDAS nodes allow access to public services, not to private sector services such as the services provided in INFINITECH, and therefore **SPeID falls out of the scope of the project**.

The three presented tools are of significant importance for banks, FinTechs, ensuring that companies or other entities are properly handling personal data or sensitive information. Personal data from insurance companies or financial institutions needs to be anonymized or pseudonymized, and users of financial services need a reliable way of authenticating their customers using government issued IDs.

1.2 Insights from other Tasks and Deliverables

The work presented in this deliverable is based on the corresponding task "T3.5 Data Governance Mechanisms, which is included in the "WP3 BigData/IoT Data Management for SHARP Services". More specifically, this deliverable is related to the following deliverables:

- **D2.6 Specifications of INFINITECH Technologies - II** exposes the fundamental steps of the process that leads to the specification of the different project technologies that constitute the building blocks utilized by the different project pilots, including Input and Output formats, functionalities and specifications about the implementation technologies (e.g. BigData/IoT platforms, AI/ML toolkits, HPC infrastructures) that will be used to realize them. Within the proposed component groups, there is one related to Security and Privacy including Data Anonymization, Digital User Onboarding System (DUOS) and Data Protection Orchestrator (DPO).

- **D2.8 Security and Regulatory Compliance Specifications - II** specifies the standard and regulatory environment that affects the INFINITECH project. An important part of the document is focused on the GDPR given its high relevance in BigData and analytics frameworks like INFINITECH. In addition, regulations such as PSD II, MiFiD II and 4AMDL are explored since they are relevant to the financial sector. This set of regulations are analysed and its applicability is assessed with respect to the INFINITECH pilot scenarios.

- **D2.14 Reference Architecture - II** presents the second version of the INFINITECH Reference Architecture (RA). The RA provides a schema to build communication workflows between the different building blocks of the INFINITECH project (as defined on D2.6 Specification of INFINITECH Technologies II), including the Data Governance Mechanisms defined in this document.

- **D3.15 Regulatory Compliance Tools - I** analyse different regulatory compliance tools and the regulatory requirements of every pilot within the INFINITECH project. The main regulatory compliance tool developed within T3.6 (namely the Data Protection Orchestrator) will communicate with the data anonymization component developed within this task.

- **D3.12 Data Governance Frameworks and Tools I** introduces the data governance mechanisms that will be developed within the INFINITECH project. A preliminary overview and design of these mechanisms is presented, including a state-of-the-art review of the relevant technologies used by each tool.

The different data governance mechanisms will be applied and validated in the pilots of "WP7 Large-Scale Pilots of SHARP Financial and Insurance Services":

- **Pilot #4 "Personalized Portfolio Management ("Why Private Banking cannot be for everyone?""** uses the Digital User Onboarding Service developed by ATOS for user enrolment and authentication.
- **Pilot #8 "Platform for Anti Money Laundering Supervision (PAMLS)"** demonstrates the implementation of the data pseudonymization mechanisms developed by JSI to pseudonymize enriched transactional data.
- **Pilot #11 "Personalized insurance products based on IoT connected vehicles"** will make use of the data anonymization tool, in particular, the prototype implementation of the location data anonymization mechanisms described in Section 3.2 to anonymize location data reported by connected cars in real time.
- **Pilot #12 "Real world data for novel health insurance products"** will use the data anonymization tool to anonymize data collected in the Healthentia platform.

1.3 Updates to the previous version

The deliverable updates the design of the data governance mechanisms presented in D3.12 "Data Governance Frameworks and Tools I" [2]. As in every software project, tools that are under development usually present changes in its design as the development process advances. This document updates the design of each of the tools to the most current version, including the different APIs, design of new components or use cases.

The design of the data governance mechanisms presented in D3.12 "Data Governance Frameworks and Tools I" [2] included the overall architecture and workflows of each of the tools. In this version, **Section 2 updates the design of each of the tools**: further details on the class structure of the pseudonymization service are included, as well as the configuration files required; the data anonymization tool's design was

updated to include a new configuration tool to help the user navigate through the complicated process of selecting the different anonymization operations and metrics. In addition, a mock-up design of the future Graphical User Interface is presented; finally, regarding the Digital User Onboarding Service, the use cases initially presented in D3.12 "Data Governance Frameworks and Tools I" [2] are updated and described in detail.

In addition, this version presents the prototype implementations of each of the tools in Section 3, including details of the REST API operations, new functionalities (such as new utility metrics or anonymization mechanisms), or the different interfaces of the digital user onboarding service.

1.4 Structure

The structure of this document is as follows. Section 1 contains the introduction of the document, the objectives and the relationship with other tasks and deliverables. Section 2 provides an updated design of the different data governance mechanisms which are developed within T3.5. Section 3 is a technical overview of the prototype implementations of each tool, together with the technical advancements from the last deliverable. Finally, Section 4 concludes the document summarizing the most important concepts.

2 Design of the data governance framework and tools

The current section is intended to update the design of the data governance framework and tools which were initially presented on D3.12 "Data Governance Framework and Tools I". On each subsection, a detailed design of the tools is provided, including the requirements and the configuration processes and user interactions.

2.1 Data pseudonymization tool

Pseudonymization is a methodology that is used in data management and de-identification, where **personal identifiable data fields in a dataset or data record are interchanged by a set of artificial identifiers or pseudonyms**. In this way, the data records become less identifiable, but they can still be used for data analysis. It needs to be noted that it is possible to restore the pseudonymized dataset to nearly its original state by using the information for re-identification of individuals (for example, translation tables or reversible hashes).

Pseudonymization is defined under General Data Protection Regulation (GDPR) [1] in Article 4(5). By this definition, **pseudonymized data cannot be attributed to a specific data subject without the use of (separate) additional information**. It must provide a methodology to remove the potential of direct or indirect identification of the subjects. A single data record might be completely unidentifiable, however, when it is placed into a larger dataset, some information might become identifiable by comparing multiple records. For example, it would be possible to identify the biggest companies by the sum of their transactions in correlation with the timestamps of transactions.

Also, GDPR Article 25(2) requires controllers to:

> *"…implement appropriate technical and organizational measures for ensuring that, by default, only personal data which are necessary for each specific purpose of the processing are processed. That obligation applies to the amount of personal data collected, the extent of their processing, the period of their storage and their accessibility. In particular, such measures shall ensure that by default personal data are not made accessible without the individual's intervention to an indefinite number of natural persons."*

The tool developed within INFINITECH will be used for pseudonymization of financial transactions' data, but the service itself will be general enough so that it will be able to handle various types of inputs. Typical data fields that need to be pseudonymized in transactional data are: names, company names, bank identificators, IBAN numbers, but also amounts, timestamps and textual data (comments, transaction descriptions, etc.).

In particular, the pseudonymizer will be **exclusively used through an API** and therefore, no user interface is envisioned for this. Pseudonymization flow will be set up with an HTTP POST request, returning a flow key. Pseudonymization process will be then performed through other HTTP POST requests, using the flow key and the data to be anonymized. **The formal API definition is provided on Section 3.**

It should also be noted, that **after the pseudonymization process the dataset cannot be further enriched**, therefore all the enriching (e. g. from other registers and datasets) should be performed before this process. As an example, after the pseudonymization of IBAN data is performed, data fusion of various connected data sources is not possible. The latter should therefore be accomplished before this step.

The pseudonymization tool consists of a hierarchy of classes as depicted in Figure 1. By using inheritance, the implementation will be more consistent and systematic. In general, three different classes of data fields are expected, where each of them will be handled by a specific pseudonymizer method: text, numeric and timestamp.

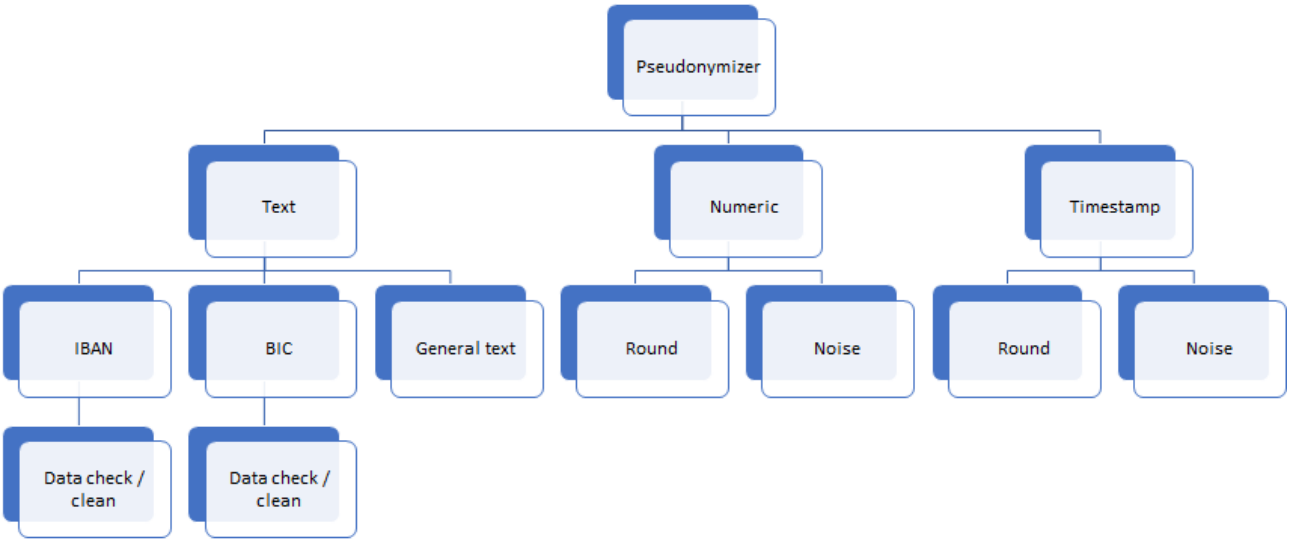


Figure 1 – Structure of classes within the pseudonymizer.

Text will be rendered in three major different ways: the simplest procedure will be used for general text, where it will be simply hashed by a selected method, whereas IBAN and BIC data will be firstly checked for consistency and cleaned (preliminary analysis of the raw data shows that there might be additional spaces in the data or inconsistent number of spaces used). For text pseudonymization, a method will be selected among hashing or encryption algorithms, such as MD5 or Blowfish. Numeric values and timestamps will not be cleaned, however, the system will enable rounding the values and even introduction of noise.

This configuration has to follow a specific structure as depicted below:

```
[
   { field: "PAYEE", type: "IBAN",
     method: "aes128", key: "key", iv: "17996d093d28ddb3ba695a2e6f58562e" },
   { field: "PAYEE_ID", type: "text",
     method: "aes128", key: "key", iv: "17996d093d28ddb3ba695a2e6f58562e" },
   { field: "PAYEE_BANK", type: "BIC",
     method: "md5" },
   { field: "DATE", type: "timestamp",
     method: "round", round_interval: 86400, noise_std: 86400 },
   { field: "SUM", type: "numeric",
     method: "round", round_interval: 1000, noise_std: 0 },
   { field: "PAYMENT_TYPE", type: "text",
     method: "none" }
   ...
]
```

The order of the defined fields is important, because in single CSV records there is no possibility to transfer the field name along with the actual record. This order is also reflected in the data that is sent to the service. We can observe that each method has its own set of parameters. Final details of the pseudonymization prototype will be given in deliverable (D3.14 "Data Governance Framework and Tools III").

A single data record can be encoded as:

```
[[ "SI56 6000 0001 1234 567", "Jožef Stefan Institute", "VNZSI22", 1623921809, 9999.90, 8, ... ]]
```

One can observe that the double bracketing (2D array) is not needed in this case, however, the same format is valid when sending multiple instances. For example:

```
[
  [ "SI56 6000 0001 1234 567", "Jožef Stefan Institute", "VNZSI22", 1623921809, 9999.90, 8, ...
],
  [ "SI56 6000 0001 1234 567", "Jožef Stefan Institute", "VNZSI22", 1623921810, 42.42, 1, ... ],
  [ "SI56 6000 0001 1234 567", "Jožef Stefan Institute", "VNZSI22", 1623921811, 3.14, 9, ... ]
]
```

As mentioned before, alternatively the data can be encoded in CSV format.

It is important to highlight that the success of a pseudonymization process of textual values is highly dependent on the input data quality. This is simply because there are large differences between very similar values after pseudonymization. A single additional space can render the pseudonymized IBAN value useless. In order to deal with this issue, BIC and IBAN cleaning strategies have been designed, according to the sample input data. These strategies eliminate certain characters or substrings from the input data. Data quality and particularities in the IBAN and BIC should be checked and addressed with the cleaning strategy prior to the pseudonymization process. The current configuration works well with the data for the BSI use case (Pilot #8 "Platform for Anti Money Laundering Supervision (PAMLS)").

Environments with smaller numbers of subjects are also sensitive to pseudonymization. In such environments, a number of more distinct entities could be recognised by their behaviour (a number of expected transactions, for example). These should be addressed by the configuration too (for example, we have implemented time and value distortion).

As it can be seen in the configuration structure above, **the pseudonymization tool will be completely configurable**, as any pseudonymization process envisioned within INFINITECH can be described with the configuration structure. The supported algorithms will be AES 128, 192, 256 (counter mode only), blowfish, sha1, sha256, sha3, md5 and ripemd160. Dedicated data cleaning will be implemented for IBAN (according to our initial validation of the sample data set). From the pre-processing also noise with rounding for numeric and timestamp values will be implemented (Gaussian with standard deviation).

2.2 Data anonymization

The aim of the anonymization is to **process personal data in order to irreversibly prevent identification**. Through this mechanism, the data can be modified in many different ways and degrees and these modifications will change the privacy and utility of the dataset. In general, as the anonymization of the data increases so does their privacy but at the expense of their utility which decreases. Thus, there exists a trade-off between these two levels which must be defined by the final user when an anonymization procedure is required. The anonymization tool that is being developed within the INFINITECH project will allow the user to select the anonymization level that best fits the required privacy and utility.

The initial design of the anonymization tool was presented on D3.12 "Data Governance Framework and Tools I" [2] and included an analysis module to select the best anonymization configuration, and an anonymization service (as a REST API) to apply the desired anonymization to a dataset. The tool required the user to provide a JSON file including a description of the data types, the different anonymization configurations to apply, and the privacy and utility metrics to calculate over the anonymized data. In the previous version this process was not user-friendly, since **the user must manually introduce the data following a specific pattern in the JSON file**. In addition, if the dataset contains different columns to which the same operation will be applied, the same information must be introduced manually multiple times.

In this phase, the tool evolved to include a **configuration module** to guide the user in the selection of the anonymization configuration, and a **Command Line Interface (CLI) to execute all the modules**

(configuration, analysis, anonymization), allowing to run the anonymization tool on environments where a graphical user interface is not available, guiding the user to provide the required information on each step.

The **configuration module** was designed to help the user set the different anonymization operations that can be performed over the dataset. Figure 2 shows an example execution of the CLI tool in configuration mode. As it is illustrated in Figure 2, the tool asks the user to provide the location of the data (1) the delimiter of the data in the CSV file (2), the guessed data types (3), asking the user for confirmation (4), the desired anonymization operations (5), and the set of privacy (6) and utility (7) metrics to be computed.

Figure 2 – example execution of the CLI Configuration Tool

At the end of the process, the generated configuration (from now on referred as *analysis configuration*) includes the following information:

- **Source data information**: type of data source (file or database), location (path or connection URL and credentials), name of the source table and columns (in case of database connection) and the data type of the columns.
- **Anonymization Configuration:** set of operations defined for each column of the dataset.
- **Metrics**: set of privacy and utility metrics defined for each column of the dataset.

Once the analysis configuration is ready, it can be used in **analysis mode**. As explained on Section 2.2. of D3.12 "Data Governance Framework and Tools I" [2] this process usually takes place over a representative sample of the data, since it is very time consuming. The objective of the analysis is to discover the set of anonymization operations that better fit the privacy and utility needs of the user.

The component first verifies the path to the data or the connection to the database where the analysis data is stored, and verifies that the operations are valid. Once the verification is completed, it computes the required privacy and utility metrics and plots the results (see Figure 3) to allow the user to make a decision on which anonymization configuration is more suitable for its needs, providing a better trade-off between privacy and utility. Once the process is finished, the resultant JSON file contains all the possible

anonymization operations that can be performed over the dataset (anonymization working point), and the values of the metrics for each operation.

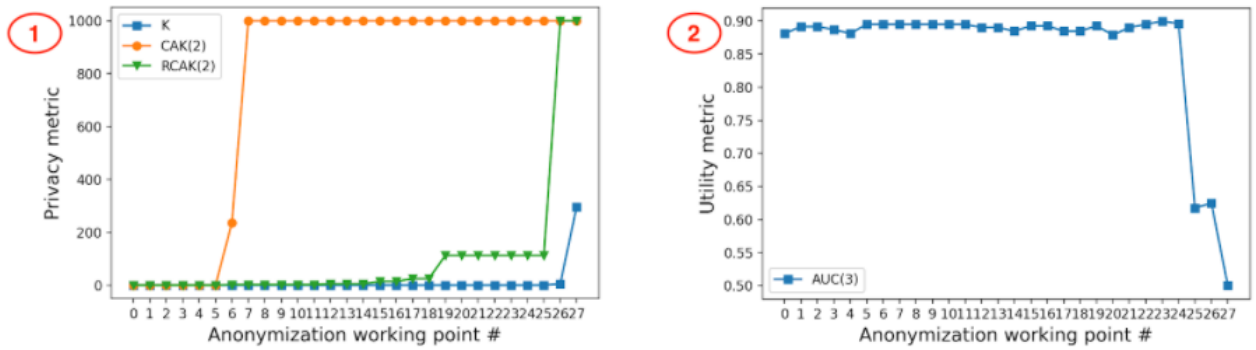


Figure 3 – metric's results for each configured anonymization operation

At this point, the user must make a decision on **whether one of the provided configurations fits its privacy and utility needs**. Figure 3 shows how the privacy and utility metrics evolve as we apply more aggressive anonymization operations (1), and how the utility evolves towards the applied anonymization (2). In this example, we can see how anonymization configuration #24 provides a good trade-off between privacy and utility, and could be a good candidate for the anonymization phase.

The user can then apply the selected configuration to the full database or **set a threshold on the value of the privacy and/or utility metric** (for example, utility metric above 0.60) and the anonymization tool will automatically select the configuration that better fits the user needs.

Once the user is ready to perform the data anonymization to the full dataset, the **anonymize mode** will ask the user for the following information, generating the *anonymization configuration*:

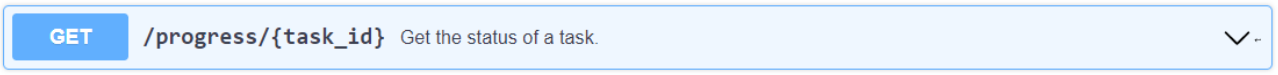
- **Parameters for the connection to the source database:** type of the database, location (URL), name of the source table, name of the columns, data type of the columns and credentials (with read-only permissions).
- **Parameters for the connection to the destination database**: type of the database, location (URL), name of the table, and credentials (in this case with write permissions).
- **Callback URL and authentication token:** since the anonymization procedure can take a long time, the operation is designed to work asynchronously. This implies that once the process has ended, the service notifies it to the client by sending a request (with the authentication token) to the specified callback URL.
- **User preferences**: preferred anonymization configuration to apply, in terms of direct anonymization configuration or specific metric thresholds.
- **Anonymization Configuration:** it includes details on the different anonymization operations that can be applied to each of the data columns. It is the result of the analysis operation.

In addition to the CLI interface to the anonymization component described above, **the tool is also intended to be used through a REST API**. The API is asynchronous to handle the long-running tasks that characterize an anonymization or analysis process. The API receives a petition to perform an analysis or anonymization operation, and starts processing the job in the background. This allows performing multiple tasks simultaneously, without the need for waiting until the operation is complete.

The endpoint */analyze* receives the analysis configuration in a JSON object, and performs the configured anonymization operations and privacy/utility metrics calculations.

If the configuration is correct, the API **returns an identifier of the task**. This identifier will allow the client to track the progress of the analysis job using the *progress* endpoint.

In a similar way, the endpoint */anonymize* receives the anonymization configuration in a JSON object. If the configuration is correct, the API returns an identifier of the task.



At last, the endpoint */progress* endpoint receives as a parameter a task identifier, and returns the current status of the task.

The implementation of the anonymization component as a service **will support the future development of a graphical user interface (GUI)**. GUIs are proven to be the most effective way to interact with the final user [3] enabling them to learn faster how to use the application. In addition, it will significantly ease the end-to-end process of configuring, analysing and applying an anonymization configuration to a database.

At this phase, the **mock-up design of the GUI was performed**, defining the interactions of the user with the anonymization tool. Figure 4 shows a sample screen of the configuration phase during the database analysis: the user will be able to add/edit different anonymization operations in an easy way, enabling a faster interaction with the anonymization component.

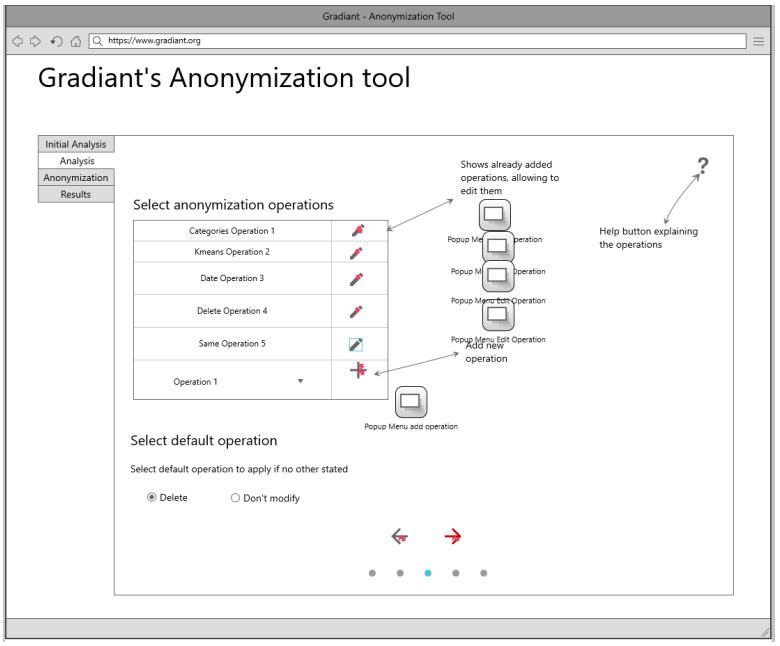


Figure 4 – mockup design of the graphical user interface

The graphical interface will allow the user to configure the anonymization operations, visualize the analysis results, and select the best anonymization configuration in an easy and visual way. The tool will also provide

detailed help sections with information about each anonymization operation to inform non-expert users about the suitability of each anonymization operation or the purpose of each privacy/utility metric.

2.3 Digital User Onboarding Tool

Digital User Onboarding System (DUOS) is a **solution for dealing with virtual identities in a mobile device**. DUOS will be used in the financial sector (including Pilot #4 "Personalized Portfolio Management ("Why Private Banking cannot be for everyone?"") of INFINITECH project) to allow bank customers or FinTechs to perform remote registration using electronic ID (eID)cards or passports. DUOS will allow them to create their own user identities (virtual eIDs) that will enable them to access the bank / fintech services but without sharing with them their biometric information. In the case of Pilot #4, DUOS will be potentially used to provide virtual identities in order to allow the access to portfolio optimization associated with this identity.

DUOS allows **remote user registration using eID or electronic passport**, providing various identity proofing and verification services in Android mobile devices. These services link the new identity (virtual eID) with a government issued eID or passport. DUOS requires the use of machine readable documents (eIDs issued by European National authorities) according to the EU eID schemas. These documents are standardized by the ICAO Document 9303 (endorsed by the International Organization for Standardization and the International Electrotechnical Commission as ISO/IEC 7501-1) [4] and have a special Machine Readable Zone (MRZ). The MRZ is specially designed to be read from electronic devices, and is usually at the bottom of the identity page at the beginning of a passport. Examples of ICAO compliant documents are eMRTD (i.e. ePassport) and the Spanish official eID card (electronic DNI) [5] .

DUOS provides multiple features to ensure the security of the created virtual eIDs:

- **Verification of the electronic data** stored on the chip.
- **Verification of the Machine Readable Zone** (MRZ) of the document using Optical Character Recognition (ORC)
- **Flexible multi-factor authentication** for different users or identities by combining face images captured from the user, with the public key certificates stored in the eID/passport.
- **Integration** with different **Identity and Access Management (IAM) systems**.

Formally, we can identify the following actors within DUOS platform:

- **User**: the user is the person interacting with the service which is requesting authorization to obtain a virtual identity and based on an official identity. The issued virtual identity is used to authenticate the user and enable access to the service. The user is the data subject providing personal data and for which the privacy protection will be implemented.
- **Verifier**. The verifier is in charge of validating the requests for authentication made by the user. The service provider grants access to its services relying on this verification procedure.
- **Service Provider** (Relying Party): provides the services the user wants to access. For example in the case of Pilot #4, the Service Provider (SP) would be the portfolio services from Prive.

DUOS supports two different use cases: **enrolment of a new user** (obtaining a new virtual ID, see Figure 5), and **authentication using an existing virtual ID** (see Figure 6). The enrolment process assumes that the eID

has already been issued by an authorized authority, and **only the enrolment of a new virtual identity (vID) falls within the scope of DUOS**. The first step is to read the official identity documents' MRZ, and extract the data from the chip. Once the data is extracted, the identity of the eID owner is verified using biometric verification using a face recognition process using a mobile device. At last, if the verification process is correct, the virtual ID is issued and stored within the mobile device to be used for authentication with third party entities (banks or FinTechs).

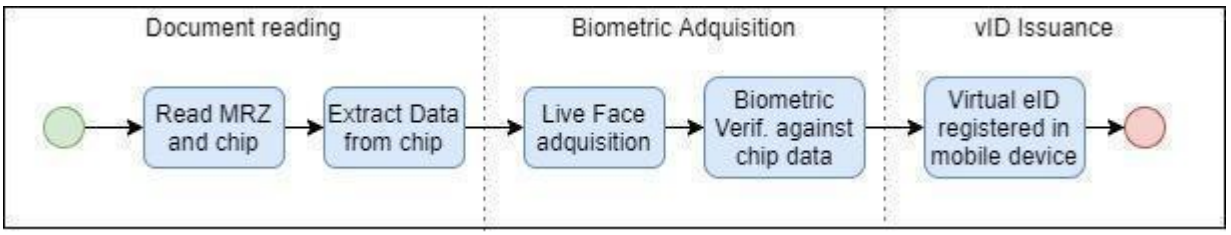


Figure 5 – DUOS enrolment use case

Once the virtual identity is issued, it can be used for authentication. The mobile device can store multiple vIDs, so the first step is to select the vID to be used. DUOS will know which third party must authenticate against by reading a QR code on the third party application.

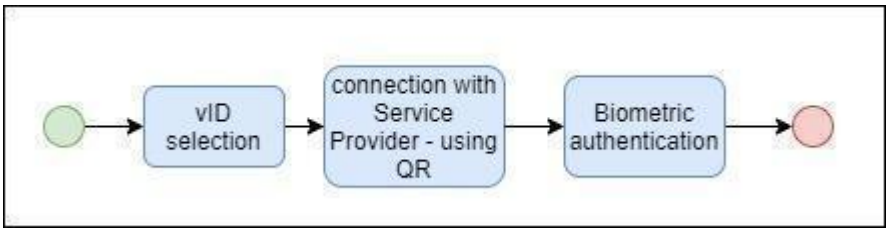


Figure 6 – DUOS authentication use case

The identity of the user is verified (to ensure that the user is who claims to be) by using **biometric authentication**. It is worth mentioning that the biometric authentication process happens within the mobile device, and the user's biometric is not shared with the third party application. In case of a successful authentication, DUOS will send to the third party the data requested by the service and perform a login action in the service provider platform.

3 Description of prototype implementation of data governance tools

This section describes the technical advances during this period for each of the tools. In each subsection a detailed technical description of the tools is provided, focusing on the advances achieved from the submission of D3.12 “Data Governance Frameworks and Tools I” [2].

3.1 Data pseudonymization service

Pseudonymization service is under development. According to the project time-line (see Figure 7) the **first version of the pseudonymization service will be delivered in October 2021**, while the final version, which will be deployed at the Bank of Slovenia (BOS) testbed in the context of Pilot #8 “Platform for Anti Money Laundering Supervision (PAMLS)” will be delivered in March 2022.

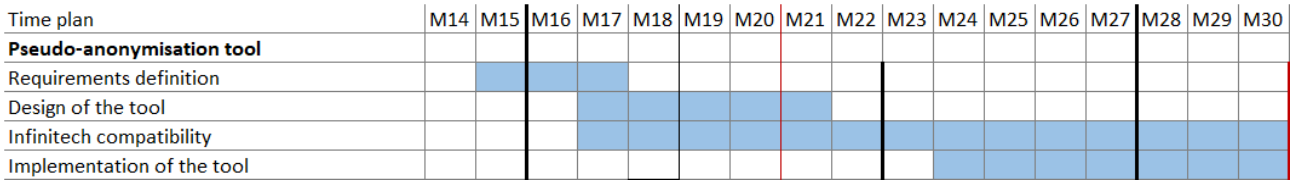


Figure 7 – detailed timeline of implementation of pseudonymization service

In this period, the **API of the pseudonymization service was defined.** The API consists of a number of different calls that will be able to configure new and existing pseudonymization flows and perform pseudonymization over individual data records, over a set of data records and even over files. The default encoding of data records will be **JSON format** (simple JSON array), however, due to potential compatibility issues we will also support processing of CSV-encoded data as well as CSV files.

Regarding authentication of the clients, in the initial version an asymmetric key pair was envisioned as an authentication mechanism. The current design makes use of **token-based authentication** which simplifies the implementation on the client side. Further description will be provided in the final deliverable for pseudonymization prototype.

This section provides an overview of the operations that can be performed using the REST API.

3.1.1 Pseudonymization flow configuration

The **first step is to configure the format of the data flows**, including each individual data field and which ones need to be anonymized. The configuration file also includes information about the pseudonymization methods to be applied to each field of the data. Each flow is assigned a *flow_id* that can be used to retrieve, delete or update an specific flow.

Table 1 – Creation of new pseudonymization flow

HTTP request	POST /flow
Description	Create a new pseudonymization flow.
Parameters	configuration *A JSON, defining configuration of the data flow, which includes parametrization of each individual data field. The latter includes the definition of the pseudonymization fields*

	(used methods and corresponding parameters). auth_token *A user-authentication token.*
Result	Results are sent in JSON format. JSON structure contains: flow_id - numeric value of the current flow id. or error - JSON array of errors, where each error is a record including parameters `code` and `description`.

Table 2 – Update of pseudonymization flow

HTTP request	PUT /flow/*flow_id* Note: *flow_id* is a parameter that identifies the pseudonymization flow.
Description	Update a specific pseudonymization flow.
Parameters	configuration *Same as with POST /flow.*
Result	*Same as with POST /flow.*

Table 3 – Deletion of pseudonymization flow

HTTP request	DELETE /flow/*flow_id* Note: *flow_id* is a parameter that identifies the pseudonymization flow.
Description	Delete a specific pseudonymization flow.
Parameters	auth_token *A user-authentication token.*
Result	*Same as with POST /flow.*

Table 4 – Get the configuracion of a pseudonymization flow

HTTP request	GET /flow/*flow_id* Note: *flow_id* is a parameter that identifies the pseudonymization flow.
Description	Get the configuration of a specific pseudonymization flow.
Parameters	auth_token *A user-authentication token.*
Result	configuration - *a configuration JSON as used in a parameter for creating a flow.*

	or error - *as specified in POST /flow.*

Table 5 – Obtain IDs of all pseudonymization flows

HTTP request	GET /flows
Description	Return ids of all pseudonymization flows.
Parameters	auth_token *A user-authentication token.*
Result	flows - *a JSON array of flow ids* *or* error - *as specified in POST /flow.*

3.1.2 Pseudonymization process

Once the flows are configured, the **pseudonymization operations can be applied** using the REST actions defined below.

Table 6 – Pseudonymization of a data record

HTTP request	POST /pseudonymize/*flow_id*
Description	Pseudonymizes a specific data record or a set of data records.
Parameters	data *Encoded in JSON, array of data records to be pseudonymized* auth_token *A user-authentication token.*
Result	data - *pseudonymized data record in JSON array* or error - *as defined in POST /flow.*

Table 7 – Pseudonymization of a data record in CSV format

HTTP request	POST /pseudonymize_csv/*flow_id*
Description	Pseudonymizes a specific data record or a set of data records in CSV format.
Parameters	data *Encoded in JSON, array of CSV data record strings to be pseudonymized*

	auth_token *A user-authentication token.*
Result	data - *pseudonymized data record in same format as input parameter* or error - *as defined in POST /flow.*

Table 8 – Pseudonymization of file in CSV format

HTTP request	POST /pseudonymize_file/*flow_id*
Description	Pseudonymizes a file in CSV format.
Parameters	file CSV file. auth_token *A user-authentication token.*
Result	file - *pseudonymized CSV file* or error - *as defined in POST /flow.*

3.2 Data anonymization

As explained in Section 2.2, the **data anonymization component is under development**. In the previous version, the analysis and anonymization modules were two different, isolated applications. The analysis tool was developed in Python while the anonymization tool was written in Java. This led to a high complexity to use the tool, since it was necessary to install and use both components separately. However, both components were aimed to be used together as a part of the same tool. With the focus on the final user and aiming to ease the use of the tool, **the application was refactored to merge and unify both tools in a single Python component**, significantly reducing the complexity of the tool. To help the user define the input, a configuration module has been implemented, as mentioned above.

Furthermore, during this period, **new utility metrics have been developed** and incorporated as a part of the tool. The objective of these metrics is to measure the remaining utility of the data after an anonymization process. The implemented metrics are based on computing the *error* or the difference between the original (non-anonymized) and anonymized data, that is, they measure the error introduced when the anonymization operations are applied. The main advantage of error-based metrics is that they can be used regardless of the final application of the data, without background knowledge of the meaning of the data points.

Table 9 shows the set of utility metrics available in the anonymization component. In particular, the Information Loss Measure (ILM) metric was recently implemented, as the other ones were already part of the anonymization tool implemented in Java.

Table 9 – Utility metrics implemented in the anonymization tool

Name	Behaviour

MSE (Mean Squared Error)	Measures the average of the squares of the errors introduced with the anonymization operations.
MAE (Mean Absolute Error)	It is an arithmetic average of the absolute errors.
MV (Mean Variation)	Measures the mean variation of the data.
ILM (Information Loss Measure)	It is an extension of Mean Error Metric, scaling the values by the square root of two times the standard deviation of the difference between both vectors.

In addition to the implementation of additional utility metrics, we continued the development of different data anonymization algorithms. A recent design was a **GPS anonymization module** based on a mechanism by Andrés *et al.* [6] to achieve geo-indistinguishability. Mathematically, given two locations x and x' and an obfuscated (anonymized) location z, geo-indistinguishability means:

$$f(z,x) \leq e^{\epsilon \times (d(x,x'))} \times f(x,\ x'),$$

where ϵ is the privacy parameter. The main idea is that, instead of disclosing the original position *x* to a location service, it is possible to add certain noise to the location, obtaining a **new anonymized position *z*** within a **radius *r* from *x***. Any other location ***x'* has the same probability of reporting the same location *z*** within a radius *r*, **which makes *x* and *x'* statistically indistinguishable**.

The geo-indistinguishability operation is integrated in the tool. **The algorithm will be demonstrated and validated in *Pilot #11 "Personalized insurance products based on IoT connected vehicles"***: Pilot #11 collects data in real time from connected cars, reporting their exact position among other data such as speed, acceleration or different vehicle related data. The collected data will be used to train different AI models to create driving profiles to derive personalized motor-insurance products based on the actual risk of a driver. In addition, real data collection from different data sources (weather related data, traffic alerts and driving behaviour) will help design a fraud detection mechanism.

In order to protect user's privacy, the location data (latitude and longitude) of the vehicle will be anonymized. Since the AI models and the post-processing of the location data assumes that the vehicle is driving in a road, **we need to remap the anonymized location to the original road where the vehicle was moving on**. This is formally known as a *remapping mechanism* [7]: an algorithm that allows the improvement of the utility of the anonymized location, while maintaining the same privacy level as the original data.

Figure 8 illustrates the design of the point-to-road remapping mechanism. Imagine a car is driving by road A: we first anonymize the position X_1 of the car (red point in the figure), obtaining an obfuscated location X_1' (green point). However, as shown in the figure, the anonymized position falls out of any road, being not appropriate for its use on a driving profile algorithm. We apply the remapping mechanism to bring the anonymized position to a point in the original road that satisfies the privacy condition given by the geo-indistinguishability condition: "*any other location x' has the same probability of reporting the same location z within a radius r, which makes x and x' statistically indistinguishable.*"

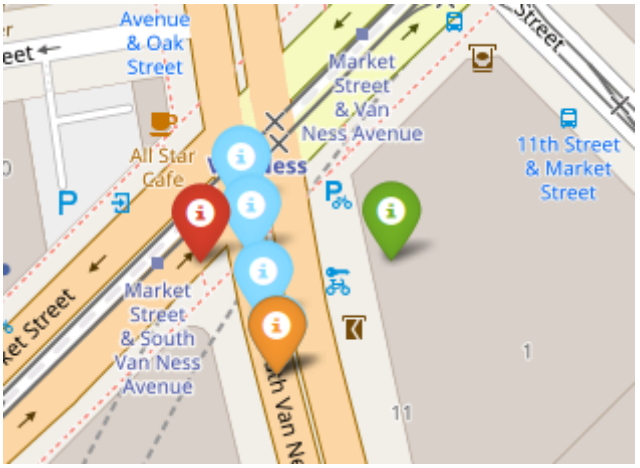


Figure 8 – point-to-road remapping mechanism

The remapping mechanism is based on finding a set of points (blue points in the figure) within a radius *r* of the original, non-anonymized one. This set of points satisfy the privacy condition and belong to the same road as the original position X_1 (utility condition). Then, we remap the anonymized point X_1' to the blue point that gives the higher privacy to the user (orange point).

However, it is worth to mention that **the point-to-road remapping mechanism is under development**, and needs to be validated with real world data in Pilot #11 "Personalized insurance products based on IoT connected vehicles". The final design of the utility-preserving GPS anonymization algorithm will be presented on D3.14 "Data Governance Frameworks and Tools III".

3.3 Digital user onboarding services tool

DUOS implementation within the INFINITECH project is an adaptation of the Digital User Onboarding System developed in [8] in the context of ARIES project. **This section summarises the required interfaces and architecture design to implement DUOS in INFINTECH**.

3.3.1 DUOS interfaces

Figure 9 showcases the DUOS interactions between the different interfaces and subcomponents. During the **enrolment phase**, the user asks to create a new identity in DUOS (a virtual entity) and provides a valid eID or passport in order to authenticate himself/herself, and DUOS extracts the user information from the eID or electronic passport (by reading the MRZ and the chip). The user is required to confirm his/her data. At last, DUOS returns the virtual identity creation result and stores it securely in the device.

Once the user has successfully created their **virtual identity** using the DUOS application, **it can be invoked by any application** (in the case of INFINITECH project, this would happen in Pilot #4 "Personalized Portfolio Management ("Why Private Banking cannot be for everyone?")" , invoked by a portal from Prive). The application requests to start an authentication process in the DUOS application, allowing the user to choose a proper virtual identity to be used in the requesting application. DUOS and the invoking application agree upon the set of required data to be provided.

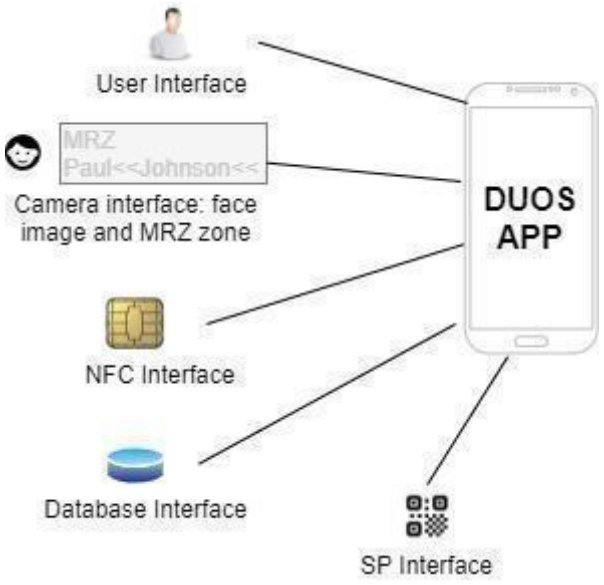


Figure 9 – DUOS interfaces

The following interfaces will be implemented to support the different features of DUOS application:

- NFC interface to read data from the eID card (e.g in Spain, eDNI Electronic National Identity Document) or electronic passport chip.
- Camera interface for (a) capturing a face image and (b) reading the MRZ of the eID card.
- SQLite database used to store the identity data of the user inside the mobile device. This interfaces uses androidx.room as abstraction layer to allow the communication with the database
- QR Reader, to allow communication with Service Providers

3.3.2 DUOS architecture

Figure 10 shows the **overall architecture of the DUOS mobile app**. This client application is mainly an eID reader and biometric verification mobile application that is used by the eID holder. As shown in Figure 10, there are two main sides where the components are deployed depending on their behaviour and functionality: the enrolment or onboarding of the user (generation of a new virtual identity), and the verification of that created vID. The verification side communicates with the service provider, returning the authentication results and the agreed data that the service provider will use.

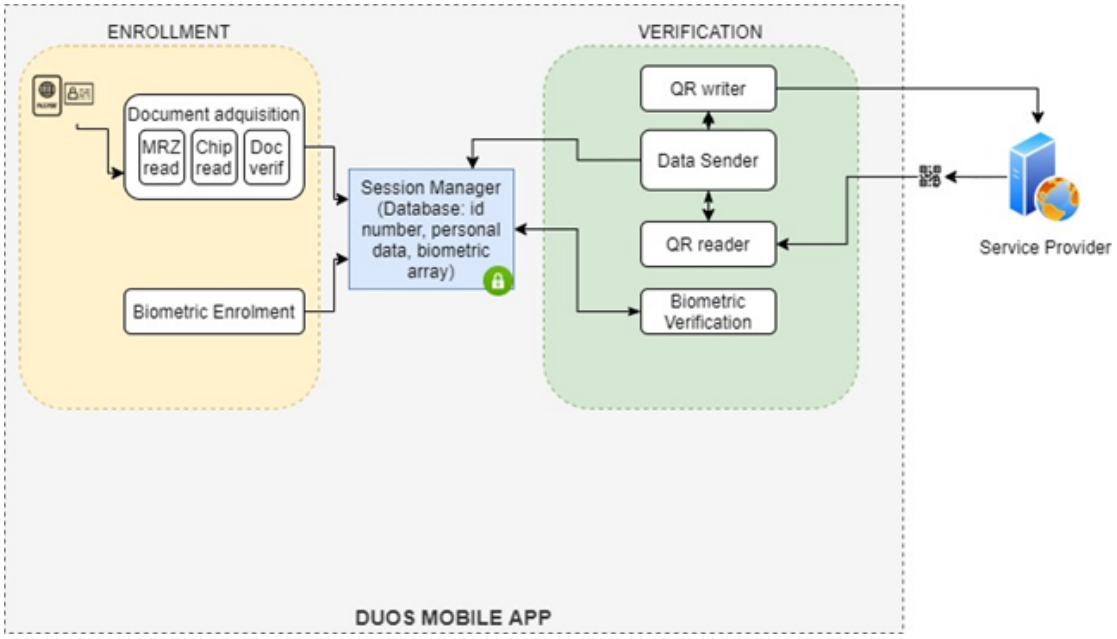


Figure 10 – DUOS mobile app architecture.

This section details the different sub-components for each module in the DUOS mobile application:

3.3.2.1 User Enrolment

This component is in charge of the **communication with electronic documents** (e-Documents) implementing the biometric verification protocol.

First of all, it is worth mentioning that **ICAO establishes a series of requirements** [9] for an inspection system (in our case DUOS) to **access the chip of an eID**: At a minimum, it should involve a process of *passive authentication.* Passive authentication consists of authenticating the digital signature inside the chip to confirm that the information that the chip contains has not been tampered. Additional mechanisms (such as active authentication or access control mechanisms) are defined by the ICAO, but DUOS is currently implementing passive authentication by reading the MRZ to extract the users' data, and then accessing the cryptographic key inside the chip to verify that the information is correct.

The **biometric enrolment module** communicates with the Session Manager, invoking the following subcomponents sequentially:

1. **MRZ reader**: together with an OCR subsystem to identify, read and translate the information from the physical document in the MRZ. If necessary, it also starts the communication with the e-Document chip. Among other information, it extracts a cryptographic key to access the chip of the e-Document.
2. **Chip reader**: this component is capable of using the NFC antenna of a mobile device to communicate with the chip of the e-Document, extracting the personal information contained on it.
3. **Document verifier**: uses passive authentication (using the key extracted from the MRZ) to access the chip of the e-Document, and obtains the required information.

The biometric enrolment module is on charge of communication with the session manager to perform all biometric-related functions:

- **Creation of biometric vectors**. Takes a face capture using the mobile device camera, and creates a biometric vector containing the extracted data.
- **Verification of biometric vectors**. Verifies that the captured biometric vector matches the biometric data stored in the e-Document chip.
- **Storage of the biometric data** in the mobile device wallet service for further access during the Virtual Identity Verification phase.

At last, the **document acquisition module** is responsible for **communication with the session manager** to provide the following functions:

- **Creation of biometric vector** from the captured biometric data using the mobile device camera to capture the face image.
- **Verify the face biometric data** captured (extracted in the previous step) against the face image stored in the chip.
- Interface with wallet to **store the biometric data** at the end of enrolment.

In this way, the **biometric data and the associated identity constitute the vID**. The Virtual ID is stored securely in the device's wallet in order to be used during the Virtual Identity Verification phase.

3.3.2.2 Virtual Identity Verification

This component is designed to establish the communication with the service provider systems, managing the **authentication of the final user in the invoking application** using the virtual ID generated during the enrolment phase.

The **biometric verification module** is responsible for:

- Capturing the face image using the mobile device camera.
- Extract the biometric vector from the device's wallet (generated during the enrolment phase).
- Verify the face biometric data captured against the face image retrieved from the vector.

As explained above, the Service Provider and DUOS agree on a set of data (information about the customer, such as name, surname, identity number, etc.) that is needed by the invoking application. The **data sender module** prepares this dataset from the virtual ID stored in the device wallet.

The **communication between DUOS and the service provider** can be established bidirectionally by either reading or writing a QR code: DUOS is capable of reading a QR code generated by the Service Provider. The Service Provider must include a URL within the code so DUOS can redirect the requested data extracted from the virtual ID. In a similar way, DUOS can generate a QR code containing the DUOS authentication results (result of the biometric verification), and the set of data that the Service Provider requested.

4 Conclusions

This document reports the progress in the scope of the task “T3.5 Data Governance Mechanisms” of the INFINITECH project, whose goal is to implement and provide the following data governance building blocks: (i) a pseudonymization tool, (ii) a mechanism for anonymizing dataset and (iii) a mobile digital user onboarding services with virtual eID derived from government issued document. Towards this goal, an updated design of the three tools was presented, updating the design documented in D3.12 “Data Governance Frameworks and Tools I” [2]. In addition, the technical advances of each of the tools during this period were described.

The pseudonymization tool that is being developed within the project will support pseudonymization of unique identifiers and generalization of numeric and time-stamp fields. It will work in batch mode by using a REST API and a configuration file provided by the user including input data attributes, the corresponding level of pseudonymization and the required type of generalization for numeric and time stamp data will be necessary. In this deliverable, the authentication mechanism was described, together with the different available pseudonymization operations, and the format of the flow configuration file required to use the component. The pseudonymization tool API is defined, but it’s technical development is not planned to start until M24 (October 2021).

Data anonymization (or de-identification) tries to handle personal data in order to irreversibly prevent identification. The analysis and anonymization configuration process is long, tedious and requires certain knowledge about the different supported anonymization operations. Toward the end of easing the use of the tool, a new configuration module was developed as a Command Line Interface tool, that helps on the definition of the anonymization operations and utility and privacy metrics. Moreover, the tool is also intended to be used through REST API: the API definition is included in the document, focusing on the required parameters, user flows and interactions with the component. Moreover, the requirement for a graphical user interface was explained, presenting the initial mock-up design of the interface. The technical advances on the anonymization component include a full refactorization of the tool to merge the two initial, isolated components, a new set of utility metrics based on the error that the anonymization introduces in the data, and a new GPS anonymization algorithm with a remapping mechanisms to keep the utility of the data in the scope of Pilot #11 “Personalized insurance products based on IoT connected vehicles”.

Paying attention to the Digital User Onboarding Service, it provides a way of enrolling new users into a service in a remote and secure way: it allows the creation of virtual identities derived from eID cards or passports, combining the electronic certificates stored in the chip with face images for increased security. The different interfaces of DUOS were described, focusing on the use cases (enrolment of new users or verification of existing virtual identities). Each of the subcomponents of the tool were described, such as the NFC interface, the QR module, or the biometric verification process. The interactions between the Service Provider and DUOS were described, focusing on the involved modules and subcomponents.

It is worth mentioning that this is the second report of the work to be done in the scope of “T3.5 Data Governance Mechanisms”, and there will be one more deliverable. Particularly, the upcoming deliverable, namely deliverable D3.14 “Data Governance Framework and Tools III”, will present the final technical design and achievements of each of the tools, and it will be delivered in M30 (March 2022). Additionally, it is expected that these tools will start being validated through the specific pilots in the following months.

Appendix A: Literature

1. European Commission. (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Da. https://eur-lex.europa.eu/eli/reg/2016/679/oj

2. "INFINITECH-D3.12 - Data Governance Framework and Tools - I." [Online]. Available: https://app.infinitech-h2020.eu/deliverable/39. [Accessed: 05-Jul-2021]

3. Feizi, Andisheh & Wong, Chui Yin. (2012). Usability of user interface styles for learning a graphical software application. 1089-1094. 10.1109/ICCISci.2012.6297188.

4. ICAO, "ICAO Document 9303 - Machine Readable Travel Documents." [Online]. Available: https://www.icao.int/publications/pages/publication.aspx?docnum=9303. [Accessed: 07-Jul-2021]

5. Machine Readable Travel Documents, Part 3: Specifications Common to all MRTDs Seventh Edition, 2015, Doc 9303, Retrieved July 2021, from https://www.icao.int/publications/Documents/9303_p3_cons_en.pdf

6. Miguel E. Andrés and Nicolás Emilio Bordenabe and Konstantinos Chatzikokolakis and Catuscia Palamidessi (2012). Geo-Indistinguishability: Differential Privacy for Location-Based Systems. CoRR, abs/1212.1984.

7. Konstantinos Chatzikokolakis, Ehab Elsalamouny, Catuscia Palamidessi. Efficient Utility Improve-ment for Location Privacy. Proceedings on Privacy Enhancing Technologies, De Gruyter Open, 2017,2017 (4), pp.308-328. 10.1515/popets-2017-0051. hal-01422842v2

8. "ARIES - ReliAble euRopean Identity EcoSystem." [Online]. Available: https://www.aries-project.eu/. [Accessed: 07-Jul-2021]

9. A-System Requirements. ICAO - Security and facilitation. Retrieved June 24, 2021, from https://www.icao.int/Security/FAL/PKD/BVRT/Pages/System-Requirements.aspx